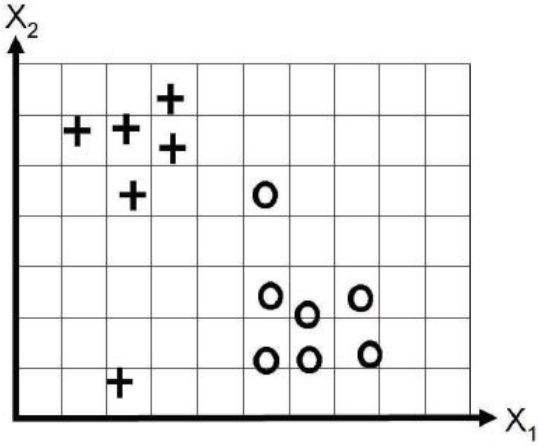
Assignment 8

Problem 1: Regularizing separate terms in 2d logistic regression



For each of the following questions, Sketch a possible decision boundary corresponding to $\hat{\mathbf{w}}$ in the

```
q1_fig.JPG
```

Be sure to answer all the questions and include the figure in your submission.

1. Consider the data in the figure above, where we fit the model $p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$. Suppose we fit the model by maximum likelihood, i.e., we minimize

$$J(\boldsymbol{w}) = -\ell(\boldsymbol{w}, D_{train})$$

where $\ell(\mathbf{w}, D_{train})$ is the log likelihood on the training set. Sketch a possible decision boundary. Is your answer (decision boundary) unique? How many classification errors does your method make on the training set?

2. Now suppose we regularize only the w_0 parameter, i.e., we minimize

$$J(\boldsymbol{w}) = -\ell(\boldsymbol{w}, D_{train}) + \lambda w_0^2$$

Suppose λ is a very large number, so we regularize w_0 all the way to 0, but all other parameters are unregularized. Sketch a possible decision boundary. How many classification errors does your method make on the training set? Hint: consider the behavior of simple linear regression, $w_0 + w_1 x_1 + w_2 x_2$ when $x_1 = x_2 = 0$.

3. Now suppose we heavily regularize only the w_1 parameter, similar to Part 2, i.e., we minimize
$$J(\boldsymbol{w}) = -\ell(\boldsymbol{w}, D_{train}) + \lambda w_1^2$$
Sketch a possible decision boundary. How many classification errors does your method make on the training set?

4. Now suppose we heavily regularize only the w_2 parameter, similar to Part 2 and Part 3. Sketch a possible decision boundary. How many classification errors does your method make on the training set?

Problem 2

Suppose we train the following binary classifiers via maximum likelihood.

- GaussI: A generative classifier, where the class conditional densities are Gaussian, with both covariance matrices set to $\mathbf{I}$ (identity matrix), i.e.,$p(\mathbf{x}|y = c) = \mathcal{N}(x|\boldsymbol{\mu_c}, \mathbf{I})$. We assume $p(y)$ is uniform.

- GaussX: as for GaussI, but the covariance matrices are unconstrained, i.e., $p(\mathbf{x}|y = c) = \mathcal{N}(x|\boldsymbol{\mu_c}, \boldsymbol{\Sigma_c})$.

- LinLog: A logistic regression model with linear features.

- QuadLog: A logistic regression model, using linear and quadratic features (i.e., polynomial basis function expansion of degree 2). After training we compute the performance of each model M on the training set as follows:
$$L(M) = \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|\boldsymbol{x_i}, \hat{\boldsymbol{\theta}}, M)$$
(Note that this is the conditional log-likelihood $p(y|\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ and not the joint log-likelihood $p(y, \boldsymbol{x}|\hat{\boldsymbol{\theta}})$ We now want to compare the performance of each model. We will write $L(M) \leq L(M')$ if model M must have lower (or equal) log likelihood (on the training set) than M', for any training set (in other words, M is worse than M', at least as far as training set logprob is concerned).

For each of the following model pairs, state whether $L(M) \leq L(M')$, $L(M) \geq L(M')$), or whether no such statement can be made (i.e., M might sometimes be better than $L(M')$ and sometimes worse); also, for each question, briefly (1-2 sentences) explain why.

1. GaussI, LinLog.

2. GaussX, QuadLog.
3. LinLog, QuadLog.
4. GaussI, QuadLog.
5. Now suppose we measure performance in terms of the average misclassification rate on the training set:

   $$R(M) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}(\boldsymbol{x_i}))$$

   Is it true that $L(M) > L(M')$ **always** implies that $R(M) < R(M')$? If so, prove it. If not, give a counter-example.

Problem 3

Please write one thing from this course so far that you found confusing, a topic you would like to hear more about, or something you found particularly interesting.